Title: Privacy Time Bombs in Omics Data

Authors: Zhiqiang Hu[1], Steven E. Brenner[1]*

Affiliations:

[1] Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA.

* Correspondence to: brenner@compbio.berkeley.edu.

The arrest of the Golden State Killer in April 2018, thirty years after a long series of murders, burglaries and rapes, was made possible because of a search of his DNA in the GEDmatch, a genealogy database that was primarily designed to identify potential relatives from direct-to-consumer genetic testing companies. Since then, many cold cases have been solved using the same approach. While this scientific advance represents a breakthrough for criminal investigations, it also poses risks to personal privacy since a significant proportion of the U.S. population could be re-identified by their DNA similarly (*1*). In addition, This also illuminates a new type of risk for research data. Fortunately, the community has foreseen the potential privacy risks of genomic data, despite not this one. Research participants' genomic data or genotypes are under controlled access in designated databases. However, privacy risks inherent in other omics data, such as RNA-seq, remain underappreciated. The raw omic reads contain genetic variants and have been regulated similarly as genomic data. High-level summary data, such as personal gene expression profiles (expression values for all genes) and DNA methylation sites, are routinely shared without restriction. Massive summary omics datasets are shared without restriction.For example, over 200,000 personal gene expression datasets generated for research participants are publicly available, e.g., through Gene Expression Omnibus (GEO) and Genomic Data Commons (GDC) Data Portal (*3, 4*).

Our recent work suggests that some types of summary omics data, including gene expression data from RNA sequencing or microarray, DNA methylation data, and DNase hypersensitive site data, pose significant privacy risks when combined with consumer genealogy databases (*2*). We found that searching a genealogy database using genotypes inferred from summary omics data, can identify a person's or their close relatives' genomic data, if existing in the genealogy database. Currently, omics data are mostly generated in research studies, usually attached with private information (e.g., disease status) and quasi-identifiers (e.g., age and sex). The combination of research participants' quasi-identifiers and their information in the genealogy database increases the re-identification risk. Once a research participant is re-identified, all information attached to their omics data, which had been deemed confidential, becomes public, violating participant consent and confidentiality. Gene expression data generated a decade ago were safe to share at the time they were created because the means to re-identify that individual

did not exist. Today, however, they do. Such risks increase over time, activated by new techniques, new knowledge, and new databases.

Here we discuss ways to mitigate the risks, including cryptographic security methods, laws and other policies aimed at addressing these new types of risk inherent in omics data while also accounting for the needs of biomedical research.

Privacy-preserving computation

One approach for mitigating concerns over privacy breaching, involves privacy-preserving strategies for sharing summary omics data. Such methods aim to preserve privacy by blocking the linkage between summary omics and genomic data while still allowing the data to be studied in biologically meaningful ways. Some privacy-preserving strategies have been applied to genomic data, including sharing aggregated data, homomorphic encryption and differential privacy (*5*).

One solution is to only share aggregated data rather than individual data. For example, for gene expression data, only the expression statistics (mean, median, or quantiles) are shared without restriction. However, this strategy will limit the data usage and their scientific impact.

In homomorphic encryption, researchers will download the encrypted summary omics data and use specifically designed tools to analyze the data. It is an appealing solution to a limited set of popular problems. However, each use case requires a specifically designed tool. The data will not be used for the analyses if no related tools have been developed.

Another strategy is differential privacy, that is, adding noise to the sensitive part of the summary omics data (*6, 7*). Adding noise to gene expression values can block their linkage to genomic data. However, our recent study suggested that a significant number of gene expression values require modification to achieve this goal. Adding more noise to a dataset makes it increasingly less useful for research purposes. Moreover, the strategy may fail if a new linking strategy is applied. Cryptographic methods are expected to degrade over time, as has been repeatedly observed in cyber security. The need to preserve individuals' genomic privacy for their lifetime and beyond while effectively sharing high-throughput molecular data poses unique challenges.

Legislation

A potentially more enduring solution for addressing privacy violations, is to pass laws against the misuse of individuals' genetic information. Many laws address discrimination, and some of those protect against genetic discrimination. For example, the U.S. Genetic Information and Nondiscrimination Act of 2008 (GINA) makes it illegal for health insurers or employers to request or require an individual's genetic information. Some U.S. states have enacted state laws against genetic discrimination in long-term care insurance, life insurance, or disability insurance. Even in the United States, laws governing DNA data are still patchy and incomplete, and new genetic privacy laws are urgently required to protect these data (*8*). While laws can reduce the level of discrimination and provide aspirational goals for society, laws alone cannot solve the

problem. If incentives exist, genomic hackers will continue to extract genetic data that reveal private information, and companies will use them to make profit.

Beyond these considerations, laws end at a jurisdiction's borders. Different cultures place different weight and value on personal privacy. Data, once available, can easily cross borders and be used legally in another country in ways entirely unintended at their point of origin. Therefore, although laws are essential to guard against genetic discrimination, they do not provide an effective solution.

Controlled access

Controlled access is a widely adopted approach. In this model, a potential user must provide a good reason to get permissions to access the data, establish they will only use them for the proposed research and provide assurances that they will keep them secure. dbGaP is one of the largest controlled access data sources, containing three million genotype data, half a million genomes and exomes, and about 165,000 transcriptomic sequencing datasets. Controlled access is an effective method for preventing data redistribution, but compared to unrestricted data, it has many limitations.

First, we estimated the citations per genome for dbGaP and 1000 Genome Project data. Since 2015, 1000 Genome Project data achieved 4 citations per genome, while dbGaP data had only <0.04 per genome, which suggests the scientific impact of a fully open genome dataset is about a hundred times that of a genome under controlled access. However, genomes under controlled access have more associated medical information. Second, obtaining data from controlled databases such as dbGaP is a time- and resource-intensive process. In addition to logistical considerations, controlled access limits the ability to carry out research to its full potential because it blocks the ability to combine data. Scientific progress is built on other people's creations. But this effort is inhibited by controlled access to data, with negative implications for discoveries that might otherwise be occurring. Ironically, the very research that people made their data available for may not be taking place. This poses dignitary harm to research participants because their goals of contributing to research are undermined.

When researchers reassure participants they will keep their data secure, they may be inadvertently misleading them. For all the restrictions put in place that impede research. the data may yet not be safe. The National Security Agency and Central Intelligence Agency may not keep their munitions secure—there are large genome data leaks as well (*9*). When people whose primary job is to keep data secure cannot ensure so, neither should we expect that from undergraduate and graduate students, whose primary job is to learn and do research.

Unrestricted sharing

At first glance, proposing the unrestricted sharing of omics data can come across like a scandalous idea. But considering we live in a world in which Google and other private companies know every place we've been, every step we take and every choice we are likely to make, we have already voluntarily given up our rights to privacy in many areas of our lives.

Consequently, it may not so unreasonable to consider a project like the Personal Genome Project (PGP), where people agree to have their data be available without restriction, acknowledging and accepting the risk that they could be re-identified, with all the potential consequences of this decision. While we cannot know in advance what harm might befall those people in the future, so are introducing risks we do not know how to quantify, maybe this situation is not so different from telling people that we will keep their data secure, when in fact we are unable to do so. One limitation is that fewer people may agree to participate in a project with unrestricted sharing. Data size of each project may be smaller, and these data as a whole may be more biased.

Omics data can introduce underappreciated privacy risks, and these risks increase over time. While no perfect solution exists, some measures must be taken to address privacy concerns. Legislation and global restrictions on genetic re-identification and discrimination are urgently needed. Although we encourage more unrestricted data availability in the future, privacy-preserving computation and controlled access should be applied, particularly for those large-scale datasets.

Acknowledgments: This work was supported by NIH grant U01 EB023686 and U41 HG0077346, and a research agreement with Tata Consultancy Services.

References and Notes:

1. Y. Erlich, T. Shor, I. Pe'er, S. Carmi, Identity inference of genomic data using long-range familial searches. *Science* **362**, 690-694 (2018).
2. Z. Hu, S. E. Brenner, Biological discovery and consumer genomics activate latent privacy risk in functional genomics data. *Submitted*, (2022).
3. A. Lachmann *et al.*, Massive mining of publicly available RNA-seq data from human and mouse. *Nature communications* **9**, 1-10 (2018).
4. L. Collado-Torres *et al.*, Reproducible RNA-seq analysis using recount2. *Nature biotechnology* **35**, 319-321 (2017).
5. A. Mohammed Yakubu, Y.-P. P. Chen, Ensuring privacy and security of genomic data and functionalities. *Briefings in bioinformatics* **21**, 511-526 (2020).
6. M. Backes *et al.*, Identifying personal DNA methylation profiles by genotype inference. *2017 IEEE Symposium on Security and Privacy*, 957-976 (2017).
7. A. Harmanci, M. Gerstein, Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nature methods* **13**, 251-256 (2016).
8. C. R. Chapman, K. S. Mehta, B. Parent, A. L. Caplan, Genetic discrimination: emerging ethical challenges in the context of advancing technology. *Journal of Law and the Biosciences*, (2019).
9. S. E. Brenner, Be prepared for the big genome leak. *Nature* **498**, 139-139 (2013).